SY09 P24 : Classification de l'état de santé du fœtus

Juliette Sacleux, Samuel Manchajm, Samuel Beziat

08/07/2024

Résumé

Ce document constitue le rapport de projet de SY09, portant sur la classification de l'état de santé de fœtus, à partir de ce jeu de données [1]. Ce rapport aborde l'exploration du jeu de données, puis la prédiction de l'état de santé d'un fœtus grâce à des méthodes de classification supervisée.

Introduction

Contexte

Depuis 2015, on assiste chaque année à 290 000 décès maternels, 1.9 million de mortinaissances (des fœtus qui meurent après 28 semaines de grossesse) et 2,3 millions de décès de nouveau-nés (au cours du premier mois de vie).

Selon les Nations Unies, la plupart de ces décès auraient pu être évités en apportant des soins appropriés. [1]. Il existe donc un réel besoin de méthodes de diagnostic efficaces et facilement accessibles.

Le *cardiotocogramme* est un examen indolore et peu couteux qui enregistre la fréquence cardiaque fœtale et l'activité utérine de la mère. Il est utilisé pendant la grossesse et pendant l'accouchement. Une meilleure interprétation de ces enregistrements permettrait une meilleure évaluation de l'état de santé des fœtus, et pourrait contribuer à réduire la mortalité infantile et maternelle.

Présentation des données

Le jeu de données est composé de 22 variables qui décrivent 2126 *cardiotocogrammes*. Ces variables concernent [2] :

— l'histogramme des rythmes cardiaques du fœtus
— l'activité utérine de la mère (contractions, mouvements du fœtus, etc.)

1. One pregnant woman or newborn dies every 7 seconds : new UN report
2. Une description plus détaillée est donnée sur Kaggle

Problématique

Chaque individu a été classé par des experts obstétriciens parmi trois catégories : **Normal, Suspect, Pathologique**, et nous supposerons ces diagnostics exacts. Notre objectif sera donc, à partir des variables, de pouvoir prédire à quelle classe un individu appartient, tout en limitant les erreurs (les faux négatifs et les faux positifs).

1 Exploration des données

Le jeu de données est exempt de valeurs manquantes, nous n'avons donc pas eu besoin d'appliquer d'imputation.

Le jeu de données est composé uniquement de variables quantitatives (continues ou discrètes), à l'exception de la variable *fetal_health*, (la variable à prédire), qui est qualitative ordinale avec 3 modalités : 1 (*Normal*) $<$ 2 (*Suspect*) $<$ 3 (*Pathological*).

Les échelles et les variances des 21 variables quantitatives sont très différentes. Nous aurons donc parfois besoin de normaliser les données (centrage et réduction).

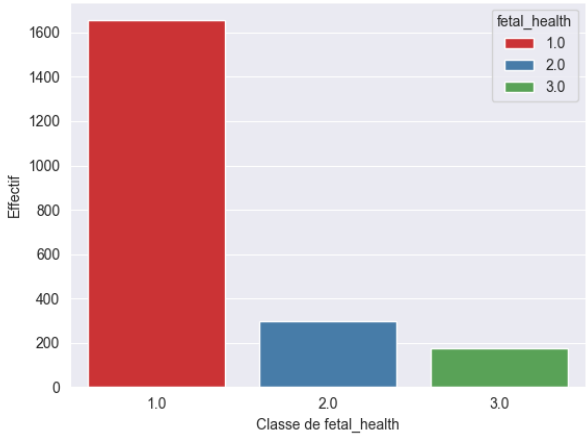

FIGURE 1 – Proportion des classes de la variable à prédire.

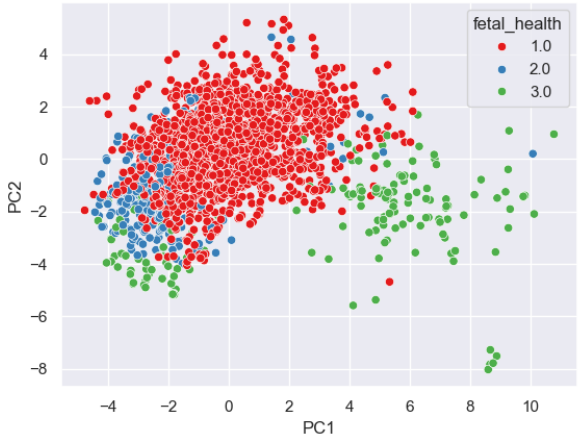

FIGURE 2 – Représentation des données dans le premier plan factoriel (ACP normalisée et pondérée)

Comme on peut l'observer sur la figure 1, le nombre d'individus dans chacune des classes est déséquilibré.

De plus, la classe minoritaire est la classe la plus "critique" (*Pathological*), il faudra donc veiller à avoir un bon score de sensibilité (peu de faux négatifs et beaucoup de vrais positifs) pour cette classe.

On observe pour certaines variables des tendances différentes en fonction des classes (c'est par exemple le cas de la variable *abnormal_short_term_variability* représentée sur la figure **??**). Nous pensons que cela pourra faciliter la tâche de classification.

Pour visualiser les données, nous avons réalisé une ACP et avons projeté les données dans le premier plan factoriel (figure 4).

Nous précisons que pour réaliser cette ACP, les données ont été normalisées pour enlever un léger "effet taille". De plus, les poids de la matrice D_p dans le calcul de l'ACP ont été modifiés, de manière à donner une importance égale à chaque classe, quel que soit son effectif :

$$p_i = \frac{1}{nb\_classes \times effectif(c(i))}$$

avec $c(i)$ la classe à laquelle appartient l'individu i.

On remarque bien que les individus sont plus ou moins regroupés spatialement selon la classe à laquelle ils appartiennent. Néanmoins, les classes ne forment pas des clusters distincts. C'est d'ailleurs pour cela que nous ne présenterons pas nos résultats pour l'algorithme des KMeans, qui ne faisait pas correspondre les classes naturelles (clusters) et les classes réelles (état de santé) (Indice de Rand Ajusté de 0.25 au maximum entre les deux partitions).

2 Sélection des variables

Pour éviter le sur-apprentissage et simplifier les modèles, nous avons créé un deuxième jeu de données où seules les variables les plus explicatives de l'état de santé du fœtus ont été conservées parmi les 22 variables disponibles.

TABLE 1 – Variables les plus liées à l'état de santé du fœtus (F score > 200.

Variable	F Score
prolongued_decelerations	505
abnormal_short_term_variability	344
%_time_abnormal_long_term_variability	345
histogram_mode	275
histogram_mean	298
histogram_median	249

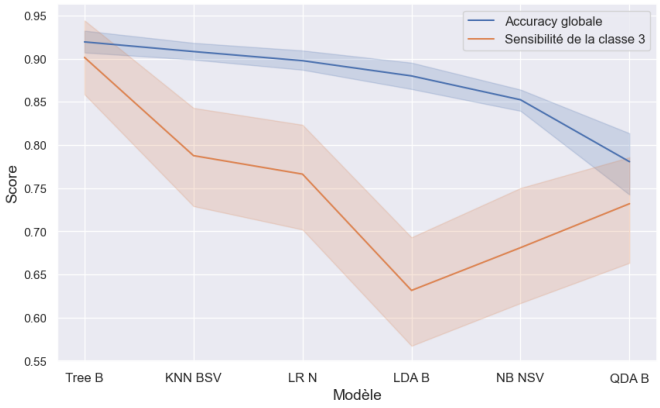

FIGURE 3 – Performances des différents classifieurs

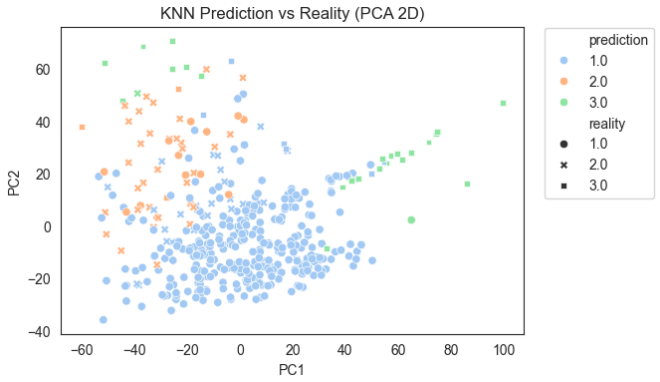

FIGURE 4 – Prédiction KNN VS Realité (représentation dans le premier plan factoriel)

Les scores du tableau 1 ont été obtenus en utilisant une ANOVA, spécifiquement avec la fonction `f_classif` du module `feature_selection` de `sklearn`. Cette méthode calcule la statistique de Fischer F entre chaque prédicteur et la variable de sortie. Plus F est élevé, plus la distribution du prédicteur est différente en fonction de la classe de l'état de santé du fœtus.

En conservant uniquement les variables avec les F-scores les plus élevés, on peut réduire la dimensionnalité du modèle et garder les prédicteurs les plus pertinents.

Par la suite, chaque méthode d'apprentissage sera entraînée avec et sans sélection des variables (nous avons fait le choix arbitraire des 6 variables les plus importantes) afin de comparer les performances. .

3 Classification supervisée

Dans cette partie, nous utilisons des méthodes de classification supervisées afin de classer nos données selon nos trois labels possibles (1 : *Normal*, 2 : *Suspect*, 3 : *Pathological*).

Pour entraîner et évaluer nos modèles, nous avons séparé nos données en 80% training et 20% test, en veillant à garder les proportions initiales de nos classes.

Nous avons recherché le modèle offrant les meilleures performances en termes de précision globale (accuracy, c'est-à-dire la proportion de bonnes prédictions) et de sensibilité (recall) pour la classe 3 (puisque les individus pathologiques sont ceux que l'on ne veut absolument pas manquer).

Pour chaque modèle, nous avons évalué les performances de 4 configurations : données brutes (B), données normalisées (N), données brutes avec sélection de variables (BSV) et données normalisées avec sélection de variables (NSV).

Pour évaluer les capacités de nos modèles de manière plus exacte, nous avons également réalisé de la cross validation avec la méthode des K-Folds (nous avons choisi K=10).

3.1 Méthode des KNN

Le premier modèle que l'on a testé est celui des K plus proches voisins.

Pour chaque configuration, nous avons calculé le nombre de voisins optimal, en effectuant une validation croisée à 10 plis sur les données d'entraînement, avec comme critère l'accuracy. Nous avons ensuite observé les performances de ces 4 modèles, en réalisant là aussi de la validation croisée à 10 plis. (figure 5).

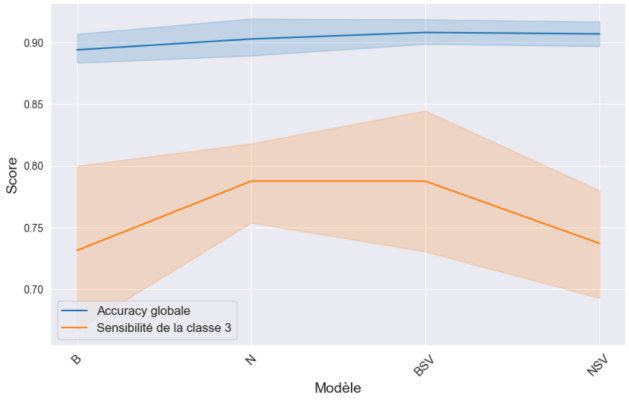

FIGURE 5 – Performances des 4 classifieurs KNN

Dans tous les cas, nous obtenons une accuracy globale sur les données de test satisfaisante (environ 0.90).

L'accuracy la moins bonne est pour B, bien que la différence ne soit pas significative.

La sélection de variables augmente la sensibilité de la classe 3, car nous supposons qu'en diminuant la dimensionnalité, les distances calculées entre les points sont plus significatives, ce qui améliore l'efficacité des KNN.

De même, la normalisation augmente les scores, probablement car elle accorde plus d'importance à des variables ayant une variance faible mais pouvant être déterminantes pour la classe 3.

Remarquons que la combinaison des deux (NSV) n'améliore pas les résultats.

TABLE 2 – Performances du classifieur KNN avec séléction de variables (moyennes des 10 plis de notre validation croisée)

Classe	Précision	Rappel	F1-score
Normal	0.94	0.96	0.95
Suspect	0.68	0.68	0.67
Pathological	0.91	0.79	0.84
Global	0.91		

Accuracy : 0.91

3.2 Analyse discriminante

À la suite des KNN, nous avons testé des modèles d'analyse discriminante : l'Analyse Discriminante Linéaire (LDA), l'Analyse Discriminante Quadratique (QDA), et le classificateur Naïf Bayésien Gaussien (NB).

Pour chacun de ces 3 classifieurs, nous avons essayé les 4 configurations décrites plus haut et avons effectué une validation croisée pour estimer les performances.

Le classifieur ayant donné les moins mauvais résultats en termes d'accuracy globale est la LDA avec données normalisées (figure **??**), pour laquelle l'accuracy était légèrement inférieure à 0.90. En revanche, la sensibilité de la LDA à la classe 3 (inférieure à 0.65 en moyenne) est moins bonne que pour les meilleurs classifieurs QDA et NB (0.68 et 0.73 en moyenne).

Dans tous les cas, la précision globale et la sensibilité à la classe 3 sont nettement moins bonnes que celles des KNN. Mais il n'est pas très étonnant que ces classifieurs ne donnent pas de très bonnes performances :

— Ces modèles ont de fortes hypothèses sur la distribution des données (normalité), or la plupart de nos variables n'ont pas des distributions gaussiennes.

— Les modèles LDA et QDA donnent des frontières respectivement linéaires et quadratiques. Puisque les classes se chevauchent, il paraît difficile de trouver de telles frontières qui séparent efficacement les classes.

— Le classifieur Naïf Bayésien Gaussien (qui donne les moins bons résultats), suppose l'indépendance des variables, ce qui n'est pas très réaliste.

3.3 Régression Logistique

Nous avons ensuite essayé de prédire la classe des individus grâce à la régression logistique. Comme pour les autres classifieurs, nous avons comparé les performances selon les 4 configurations de données. Nous avons utilisé la validation croisée pour estimer les performances (figure 6).

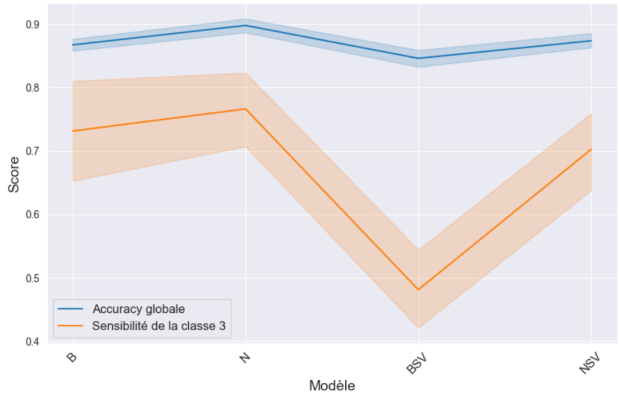

FIGURE 6 – Performances des 4 modèles de Régression Logistique

On remarque immédiatement que, contrairement aux KNN, la régression logistique a donné de meilleurs résultats sans sélection de variables.

En effet, la sélection de variables était utile pour KNN, car elle évitait le fléau de la dimension, et rendait donc les distances plus significatives. Or la régression logistique n'utilise pas les distances entre les points, donc le fléau de la dimension pose moins de problèmes. Supprimer des variables ne fait que supprimer de l'information.

La régression logistique sur données normalisées (N) a donné une accuracy globale légèrement meilleure à la régression logistique sur données brutes (B). De la même manière, la sensibilité à la classe 3 est meilleure pour les données normalisées. Ces différences s'expliquent par la nécessité de normaliser les données pour assurer la convergence de l'algorithme.

Les résultats sont assez proches des résultats des

KNN.

TABLE 3 – Performances de la régression logistique avec normalisation des données (moyennes des 10 plis de notre validation croisée)

Classe	Précision	Rappel	F1-score
Normal	0.95	0.95	0.95
Suspect	0.67	0.67	0.66
Pathological	0.85	0.77	0.80
Global	0.90		

Accuracy : 0.90

3.4 Arbres de décision

Enfin, nous avons utilisé les arbres de décision binaires pour classifier nos données. C'est le modèle qui a donné les meilleurs résultats. Comme pour les autres modèles, nous avons essayé les 4 configurations de données et avons utilisé la validation croisée pour estimer les performances de chaque modèle (figure 7).

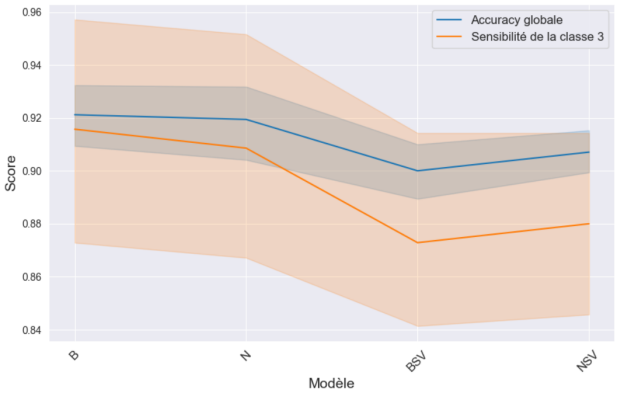

FIGURE 7 – Performances des 4 modèles d'Arbre de Décision

Comme la régression logistique et contrairement au KNN, nos arbres de décision fonctionnent mieux sans sélection de variables. C'est tout à fait logique puisque les arbres de décision sélectionnent les variables les plus pertinentes à chaque séparation. Il n'y a pas besoin de sélectionner des variables pertinentes à l'avance, cela ne fait que supprimer de l'information.

L'arbre de décision sans normalisation a une meilleure précision globale que celui avec normalisation, et une meilleure sensibilité à la classe 3. Toutefois, les scores des 2 modèles sont très proches, et surtout, ils sont tous les deux meilleurs que l'ensemble des modèles que nous avons essayés (nous avions, au mieux, une accuracy de 0.90 et une sensibilité à la classe 3 inférieure à 0.80). Ainsi, le gain en accuracy globale est assez léger par rapport aux KNN (0.92 au lieu de 0.90). En revanche, le gain en sensibilité pour la classe 3 est important (≈ 0.90 au lieu de 0.79).

TABLE 4 – Performances de l'arbre de décision avec les données brutes (moyennes des 10 plis de notre validation croisée)

Classe	Précision	Rappel	F1-score
Normal	0.96	0.95	0.96
Suspect	0.75	0.76	0.76
Pathological	0.91	0.91	0.91
Global	0.92		

Accuracy : 0.92

Nous pouvons tenter d'expliquer pourquoi ce modèle fonctionne mieux que les précédents :

- Là où KNN souffre du fléau de la dimension, les arbres de décision ne nécessitent pas de calcul de distances et peuvent mieux gérer des espaces de haute dimension en découpant l'espace de manière récursive.
- Là où LDA et QDA posent de fortes hypothèses sur les données, les arbres de décision ne font aucune supposition sur celles-ci.
- Là où le classifieur Naïf Bayésien Gaussien suppose l'indépendance conditionnelle des variables par rapport à la classe cible, les arbres de décision capturent les interactions entre les variables sans supposer leur indépendance.
- Là où la régression logistique se limite à modéliser une relation linéaire entre les caractéristiques et les log-odds de la variable cible, les arbres de décision capturent des relations non linéaires complexes.

4 Conclusion

Le problème sur lequel nous travaillons consiste à prédire l'état de santé du fœtus à partir de 21 variables quantitatives issues de *cardiotocogrammes*.

4.1 Principaux résultats

Nous avons essayé plusieurs classifieurs, avec différents prétraitements de données. La figure 3 montre les meilleures performances que nous avons obtenues pour chaque classifieur.

Parmi tous les modèles que nous avons testés, aucun n'est très mauvais, même en termes de sensibilité à la classe 3 (minoritaire).

L'arbre de décision montre les meilleurs résultats, que ce soit en termes de précision globale (0.92), ou en termes de sensibilité à la classe 3 (0.91). Ce modèle se démarque réellement des autres pour l'identification des cas pathologiques, ce qui est très important dans une application critique comme celle étudiée.

4.2 Perspectives

Nous avons réussi à obtenir de bons résultats en termes d'accuracy et de sensibilité à la classe *Pathological*, en particulier grâce aux arbres de décision.

Néanmoins, il reste toujours des pistes pour améliorer nos résultats. Par exemple, certaines techniques telles que l'utilisation du suréchantillonage ou du sous-échantillongage, ou bien l'introduction de coûts associés aux décisions pourraient améliorer encore les sensibilités de nos différents modèles à la classe minoritaire.

De plus, on peut s'interroger sur les critères que nous avons choisis pour comparer nos modèles. Nous avons considéré que les scores à maximiser étaient l'accuracy globale et la sensibilité aux *Pathological*, mais ce choix est discutable. En effet, si un fœtus appartenant à la classe 3 est classifié comme *Suspect*, on peut imaginer qu'il bénéficiera tout de même d'une surveillance, et cela serait ainsi moins grave que s'il avait été classé comme *Normal*. Nous aurions ainsi pu nous intéresser à la proportion de *Pathological* classés comme *Suspect* ou *Pathological*, plutôt qu'à la simple proportion de *Pathological* classés comme *Pathological*.

Références

[1] D. A. de Campos, J. Bernardes, A. Garrido, J. M. de Sá, and L. Pereira-Leite. Sisporto 2.0 : A program for automated analysis of cardiotocograms. Technical report, 2000.

5 Annexes

TABLE 5 – Performance des algorithmes de classification

Classifier	Feature Selection	Standardization	Acc_train	Acc_test	Recall_Path
Knn	False	False	0.918824	0.884977	0.742857
Knn	True	False	0.947059	0.901408	0.742857
LDA	False	False	0.888235	0.847418	0.542857
LDA	True	False	0.860588	0.823944	0.485714
LDA	False	True	0.888235	0.847418	0.542857
LDA	True	True	0.860588	0.823944	0.485714
QDA	False	False	0.730000	0.720657	0.628571
QDA	True	False	0.844118	0.802817	0.571429
QDA	False	True	0.636471	0.638498	0.600000
QDA	True	True	0.844118	0.802817	0.571429
GaussianNB	False	False	0.821176	0.800469	0.600000
GaussianNB	True	False	0.855294	0.833333	0.628571
GaussianNB	False	True	0.734118	0.713615	0.600000
GaussianNB	True	True	0.852353	0.814554	0.628571
Logistic Regression	False	False	0.890000	0.870892	0.800000
Logistic Regression	False	True	0.907059	0.875587	0.657143
Logistic Regression	True	True	0.874118	0.842723	0.685714
Decision Tree	True	False	0.999412	0.920188	0.885714
Decision Tree	False	False	1.000000	0.931925	0.885714
Decision Tree	True	True	0.999412	0.922535	0.914286
Decision Tree	False	True	1.000000	0.931925	0.914286